TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

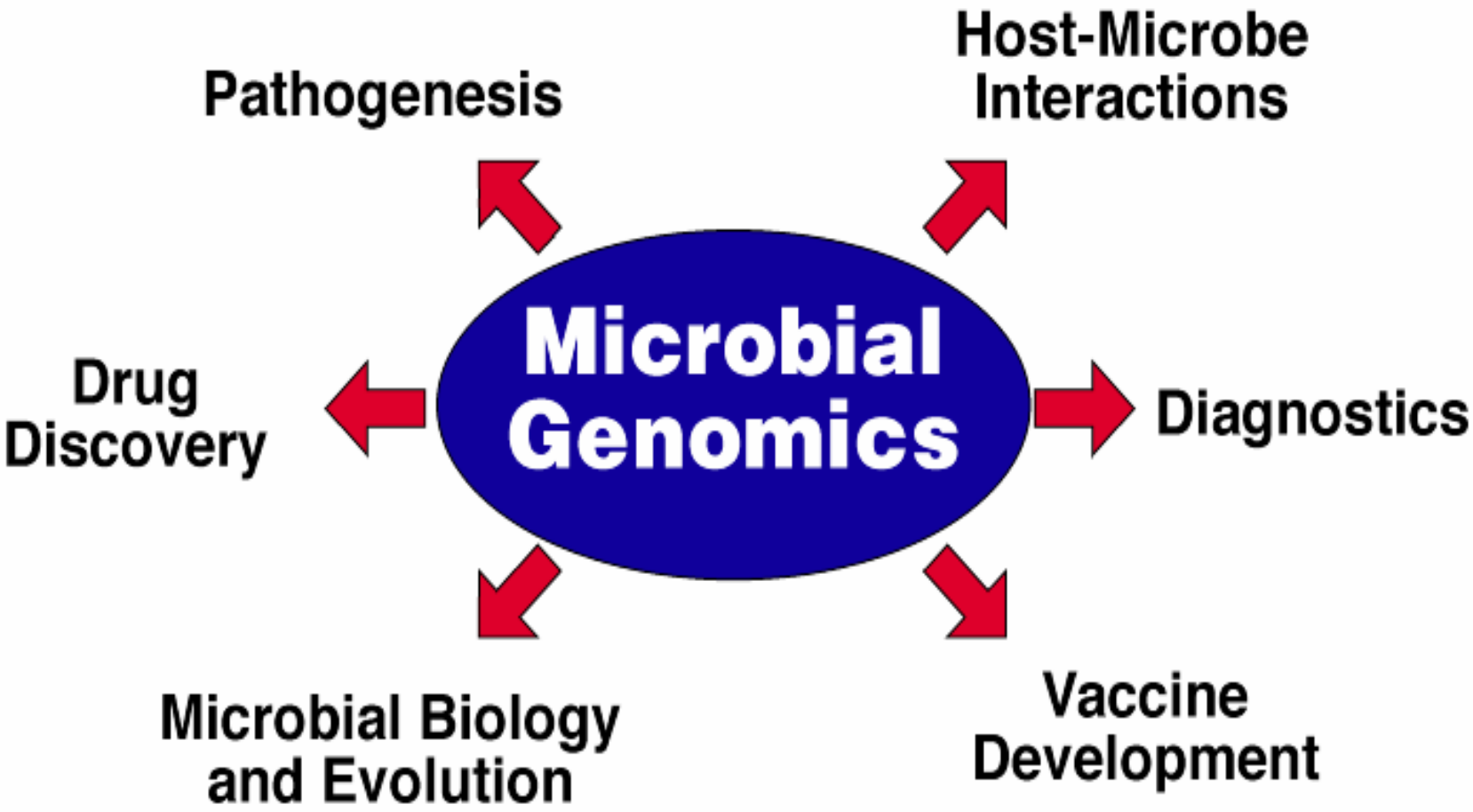
TIGR's MSC Program

Dr. Eric Eisenstadt, Ph.D.
VP for Research

Topics

(not necessarily in this order)

- MSC Overview and Mandate
- Genomes and Workflow across Resource Centers
- Data Release Policy & Reporting
- New Project Evaluations
- Current MSC Status & Completion
- Program Interactions


Pathogenesis
Host-Microbe Interactions
Drug Discovery
Microbial Genomics
Diagnostics
Microbial Biology and Evolution
Vaccine Development

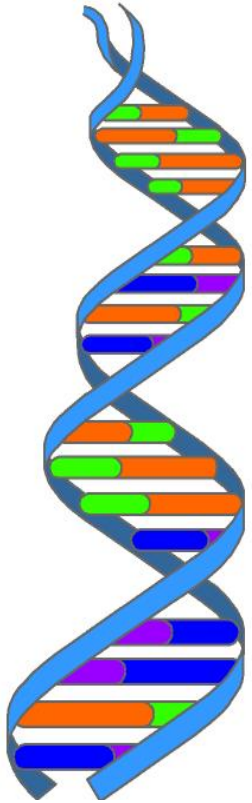
NIAID Microbial Genome Sequencing Centers

Goal:

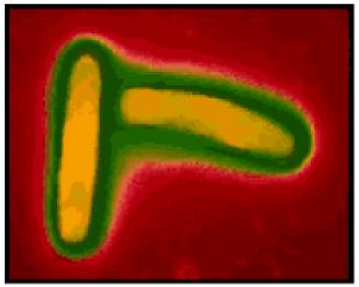
Rapid and cost-efficient production of high-quality genome sequences of human pathogens and invertebrate vectors of disease.

Features:

- Two Genome Sequencing Centers
  - The Institute for Genomic Research (TIGR)
  - MIT
- Capacity to sequence genomes for:
  - Other government agencies
  - Scientific community
  - Response to national emergencies



Mining Microbial Genomes to Develop Countermeasures



Pathogen

Cloned Genomic Sequence

Mechanisms of Pathogenesis

Virulence Factors

Structural Proteins

Metabolic Pathways

Immune Evasion Factors

Biosynthetic Pathways

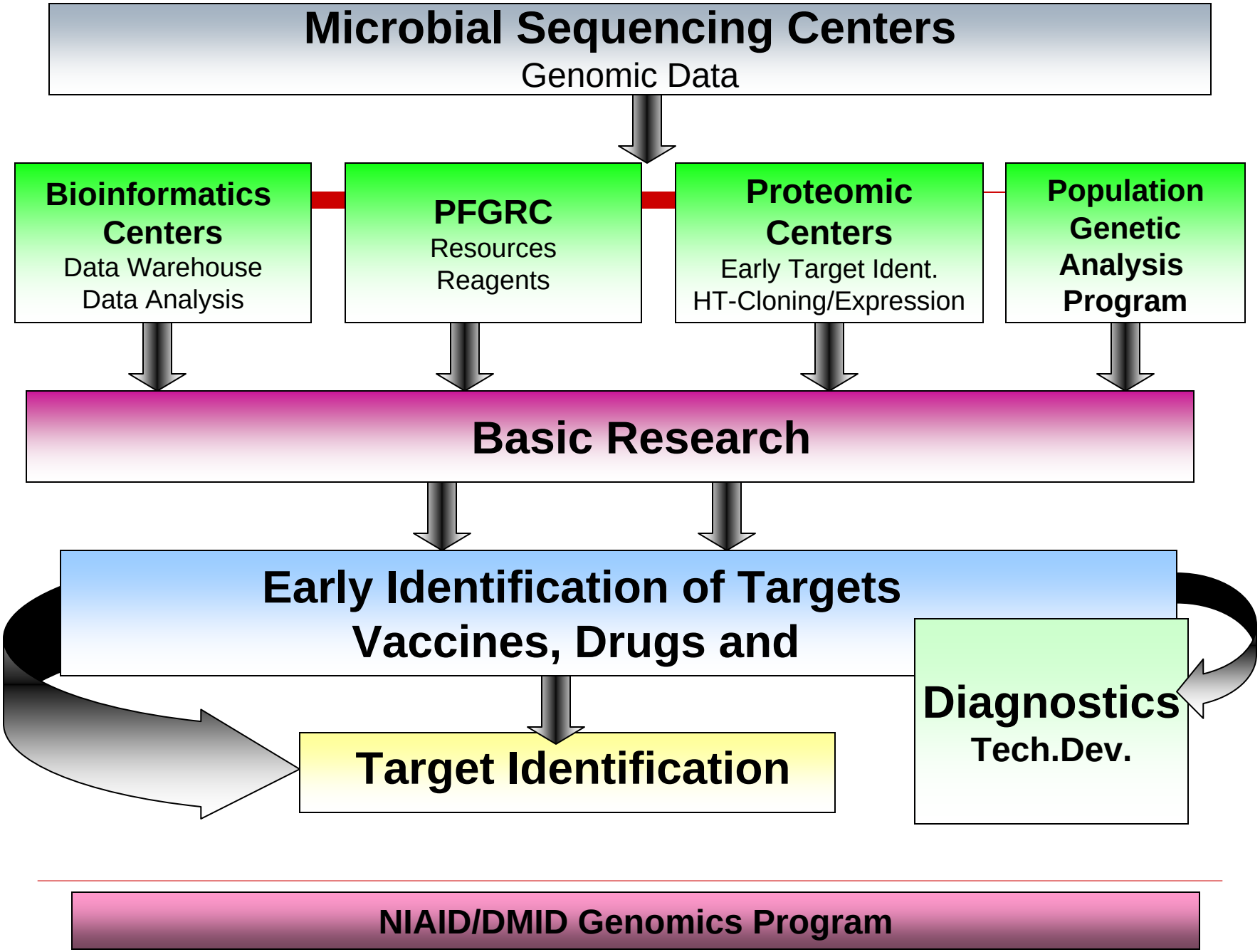
Drug-resistance Mechanisms

Receptors

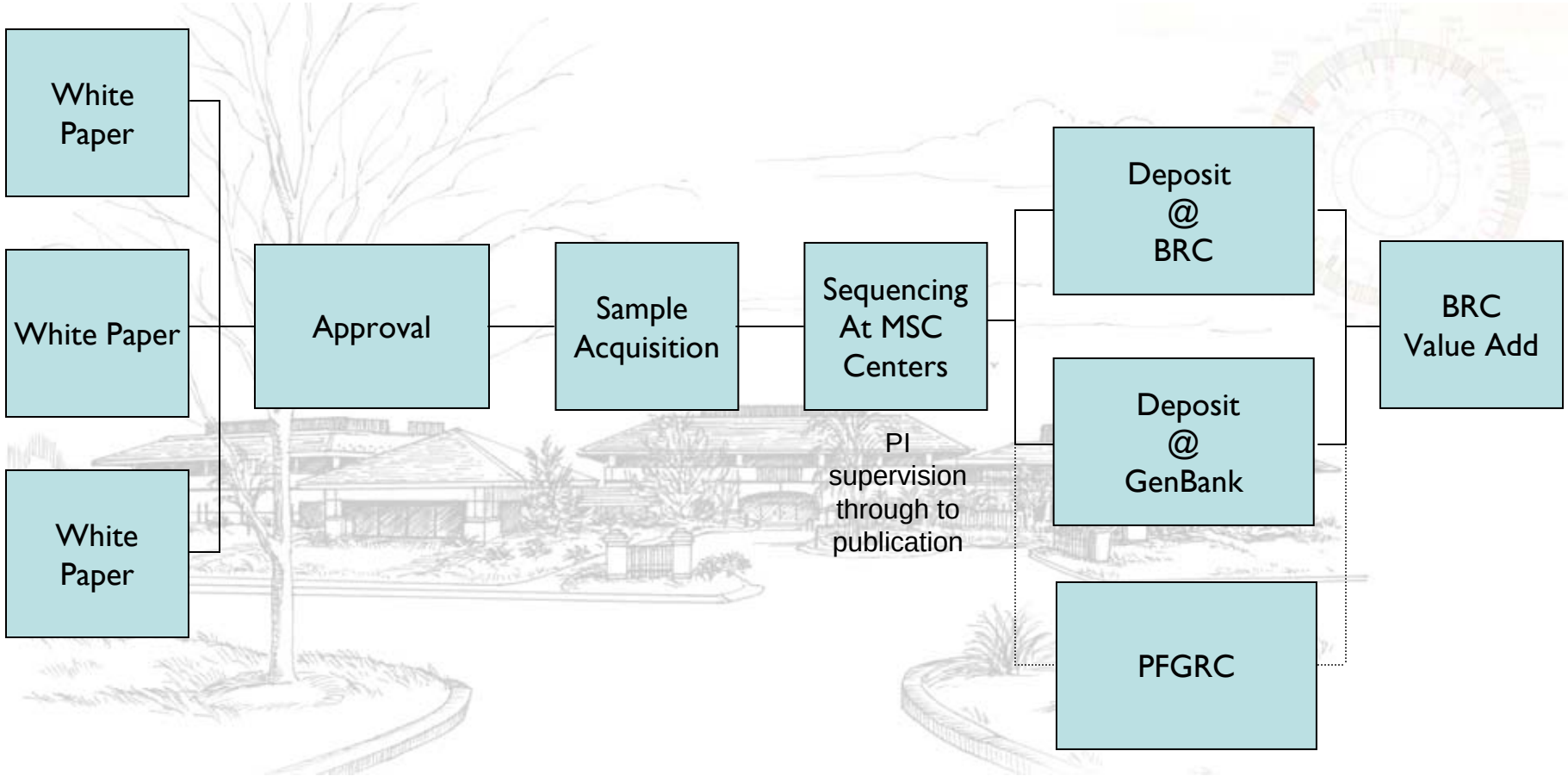
Unique Sequences (Rapid Diagnosis)

Immunostimulatory Proteins

Whole Genomic Sequence

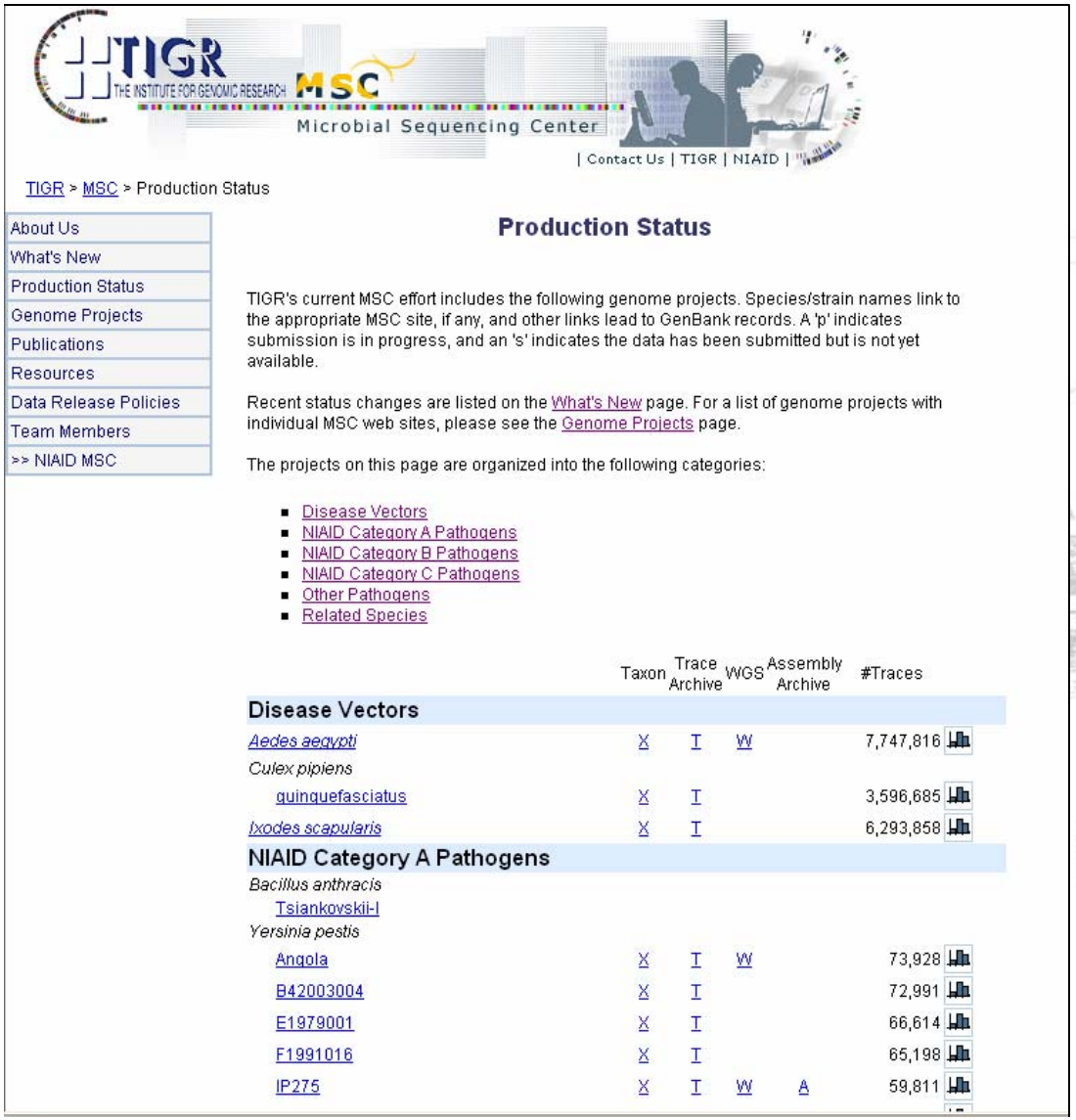

Microbial Sequencing Centers
Genomic Data
Bioinformatics Centers
Data Warehouse
Data Analysis
PFGRC
Resources
Reagents
Proteomic Centers
Early Target Ident.
HT-Cloning/Expression
Population Genetic Analysis Program
Basic Research
Early Identification of Targets
Vaccines, Drugs and
Diagnostics
Tech.Dev.
Target Identification
NIAID/DMID Genomics Program

Genome Work and Data Flow Across the TIGR/NIAID Resource Centers



Data Release & Reporting

http://msc.tigr.org/status.shtml



TIGR THE INSTITUTE FOR GENOMIC RESEARCH MSC Microbial Sequencing Center

| Contact Us | TIGR | NIAID |

TIGR > MSC > Production Status

- About Us
- What's New
- Production Status
- Genome Projects
- Publications
- Resources
- Data Release Policies
- Team Members
- >> NIAID MSC

Production Status

TIGR's current MSC effort includes the following genome projects. Species/strain names link to the appropriate MSC site, if any, and other links lead to GenBank records. A 'p' indicates submission is in progress, and an 's' indicates the data has been submitted but is not yet available.

Recent status changes are listed on the What's New page. For a list of genome projects with individual MSC web sites, please see the Genome Projects page.

The projects on this page are organized into the following categories:

- Disease Vectors
- NIAID Category A Pathogens
- NIAID Category B Pathogens
- NIAID Category C Pathogens
- Other Pathogens
- Related Species

Taxon	Trace Archive	WGS	Assembly Archive	#Traces
Disease Vectors				
Aedes aegypti	X	T	W	7,747,816
Culex pipiens quinquefasciatus	X	T		3,596,685
Ixodes scapularis	X	T		6,293,858
NIAID Category A Pathogens				
Bacillus anthracis Tsiankovskii-I				
Yersinia pestis Angola	X	T	W	73,928
B42003004	X	T		72,991
E1979001	X	T		66,614
F1991016	X	T		65,198
IP275	X	T	W A	59,811

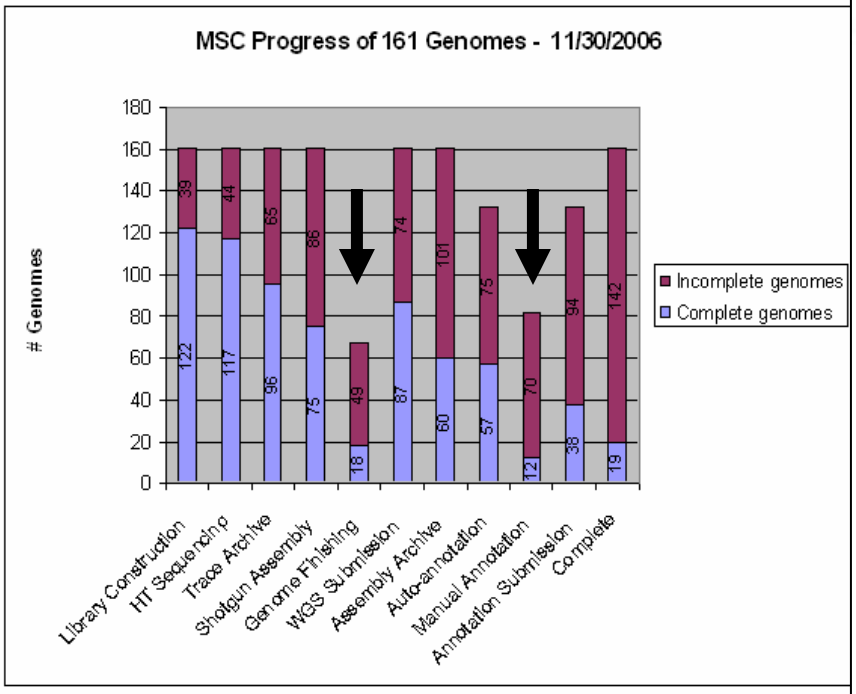
Submissions

- *Chromatograms:* daily or regular intervals based on project
- *Assembly files:* >= 3X assembly, 45 days after completion
- *Auto-annotation:* >= 5X immediately after completion
- *Manual Annotation:* After sequencing/finishing is completed

Current Status & Completion

#	ID	Name	Strains	Coverage / Closure	Auto Annotation	Manual Annotation
		Insect Vectors of Disease				
1	8010B	*Aedes*	1	n(8X)	y	n
2	8010T	*Culex pipiens*	1	n(4X)	y	n
3	8010O	*Ixodes scapularis*	1	n(3X)	y	n
4	8010N	*R. communis*	1	n(4X)	n	n
		Parasites				
5	8010Z	*Cryptosporidum muris*	1	n(8X)	y	y
6	8010M	*Entamoeba*	3	y(1)	y	y
7	8010F	*Plasmodium vivax*	1	y 9Mb	y	y
8	8010G	*Trichomonas Genome*	1	n(6X)	y	n
9	8010P	*Toxoplasma. Gondii*	2	n	y	y
		Virus				
10	8010E	RNA Lab and Influenza	5400	y	n	n
11	8010K	Coronavirus Expansion	105	y	n	n
		Fungi				
12	8010J	Comparative Aspergillus project	2	n	y	y
13	8010V	*Penicillium* spp	2	n	n	n
		Bacteria				
14	8010R	*B. cereus*	10	y(5) 27.5 Mb	y	y
15	8010S	*Bartonella baciliformis*	1	y	y	y
16	8010aa	*Borrelia*	17	y(17) 0.69Mb/genome	y	y
17	8010A	*Burkholderia mallei*	11	y(2)	n	n
18	8010D	*Burkholderia pseudomallei*	9	y(3) 14.6Mb	n	n
19	8010Q	*Campylobacter*	10	y(5) 9Mb	y	n
20	8010L	*Coxiella*	6	y(6) 10Mb	y	y(2)
21	8010H	Pathogenic *E. coli* and *Shigella*	11	y(4)	y	n
22	8010ab	*E.coli* SECEC SMS3-5	1	y	y	y
23	8010U	*Pseudomonas aeruginosa PA7*	1	y	n	n
24	8010Y	*Salmonella* Strains	17	y(9) 44.1Mb	y	y(3)
25	8010X	*Ureaplasma*	20	y(2) 2Mb	y	y(2)
26	8010C	*Vibrio* strains	17	y(1)	y	y(1)
27	8010I	*Yersinia pestis*	9	y(3) 9.2 Mb	y	n
28	8010W	454 Data Assembly R&D	n/a	n/a	n/a	n/a

Contract Milestone Deliverables in the Aggregate	Due Date	% Completed	Deliverable Status
• Sequence submissions	Sept 2008	60%	On schedule
• WGS submissions	Sept 2008	54%	On schedule
• Closure and WGS updates	Sept 2008	27%	Behind schedule
• Annotation submissions	Sept 2008	29%	Behind schedule



Note: This illustration excludes viral genomes

New MSC Projects

Organism	Strains	Relevant BRC
Methicillin-resistant *Staphylococcus* plasmids	100 plasmids	NMPDR
Vibrio parahaemolyticus	7strains	NMPDR
Vibrio cholerae – ICE	1 strain	NMPDR
Escherichia coli O157:H7	8 strains	ERIC
Neisseria meningitidis	Multiple genes	Not applicable
Murine coronavirus	10 strains	PATRIC

Selection Criteria

- Medical importance
- Status of other projects on the same organism
- Lab strain vs. clinical isolate
- Availability of experimental tools
- Community of users
- How well the strain represents the species

Sample Acquisition

- DNA / RNA
- Cell pellet

Genomes Common to NIAID/TIGR Resource Centers

#	MSC Genomes	Strains	BRC Presence
	Insect Vectors of Disease		
1	Aedes	1	VectorBase
2	*Culex pipiens*	1	VectorBase
3	Ixodes scapularis	1	VectorBase
4	*R. communis*	1	
	Parasites		
5	*Cryptosporidium muris*	1	ApiDB
6	Entamoeba	3	Pathema (E.histolytica)
7	*Plasmodium vivax*	1	ApiDB
8	*Trichomonas Genome*	1	
9	*Toxoplasma. Gondii*	2	
	Virus		
10	RNA Lab and Influenza	5400	BioHealthBase
11	Coronavirus Expansion	105	Patric
	Fungi		
12	Comparative Aspergillus project	2	
13	Penicillium spp	2	
	Bacteria		
14	B. cereus	10	
15	*Bartonella baciliformis*	1	
16	Borrelia	17	
17	Burkholderia mallei	11	Pathema
18	Burkholderia pseudomallei	9	Pathema
19	Campylobacter	10	NMPDR (*C.jejuni*)
20	*Coxiella*	6	Patric (*C.burnetti*)
21	Pathogenic E. coli and Shigella	11	Eric
22	*E.coli* SECEC SMS3-5	1	
23	*Pseudomonas aeruginosa PA7*	1	
24	Salmonella Strains	17	
25	*Ureaplasma*	20	
26	*Vibrio* strains	17	
27	*Yersinia pestis*	9	Eric

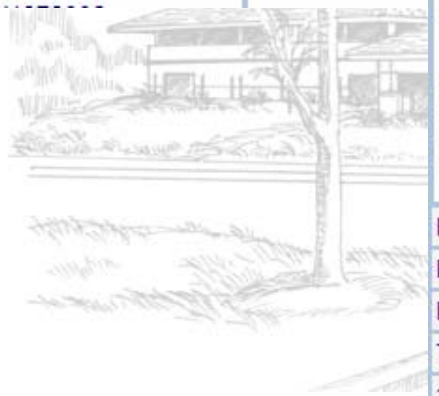
BRC Center	BRC Genome	MSC presence	PFGRC Arrays
ApiDB	*Cryptosporidium spp.*	✓	
	Plasmodium spp.	✓	✓
	Toxoplasma gondii	✓	
BioHealthBase	*Francisella tularensis*		✓
	Giardia lamblia		✓
	Microsporidia		
	Ricinus communis	✓	
	Mycobacterium tuberculosis		✓
	Influenza Virus	✓	
ERIC	*Yersinia pestis*	✓	✓
	Diarrheagenic E. coli	✓	
	Yersinia enterocolitica		
	Shigella,	✓	
	Salmonella	✓	✓
NMPDR	*Staphylococcus aureus*		
	Pathogenic Vibiros	✓	✓
	Listeria monocytogenes		✓
	Campylobacter jejuni	✓	✓
	Streptococcus pyogenes		
	Streptococcus pneumoniae		✓
Pathema	*Bacillus anthracis*		✓
	Burkholderia mallei	✓	✓
	Burkholderia pseudomallei	✓	✓
	Clostridium perfringens	✓	
	Entamoeba histolytica	✓	
Patric	*Rickettsiae*		✓
	Brucella		
	Coxiella burnetii	✓	
	Calicivirus		
	Hepatitis A Virus		
	Rabies Virus		
	Coronavirus	✓	✓
VBRC	*Variola major Virus*		
	Arenavirus		
	Hanta Virus,		
	Rift Valley Fever Virus		
	Ebola Virus,		
	Marburg Virus		
	Dengue Virus		
	California encephalitis group Virus		
	Kyansanar forest disease Virus		
	Omsk hemorrhagic fever Virus		
	West Nile Virus		
	Alphavirus		
	Hantaan Virus,		
	Puumala Virus,		
	Crimean-Congo hemorrhagic fever Virus,		
	Yellow fever Virus,		
	Tick-borne Encephalitis		
	Nipah Virus		
	Equine morbillivirus		
VectorBase	*Anopheles gambiae*		
	Aedes aegypti	✓	
	Anopheles gunestus		
	Culex pipiens	✓	
	Ixodes scapularis	✓	

MSC Program ⟷ **BRC Genomes** ⟷ **PFGRC Arrays**

About Us
What's New
Production Status
Genome Projects
Yersinia
- Goals
- Isolate Selection
- Strains
- GenBank Data
- References

Yersinia pestis
- Angola
- B42003004
- E1979001
- F1991016
- IP275

Y. pestis str
rhamnose a
Yersinia pse
Y. pestis str
pseudotube
pestis strain
typical *Y. pe*
Yersinia, the
pestis. Angol
deletion affe



About Us
What's New
Production Status
Genome Projects
A. aegypti
- Released Data
- GenBank Submissions
- BAC Ends
- TIGR Gene Index
- Sequencing Information
- GenBank Data
- Progress by Library
- Timeline [PDF]
- Project Plan [PDF]

Publications
Resources
Data Release Policies
Team Members
>> NIAID MSC

Released Data

Updated 6/14/06

***Aedes aegypti* annotation Release 1.0**

The NIAID Microbial Sequencing Centers are pleased to announce annotation Release 1.0 of the Aedes aegypti genome sequence. This annotation was produced jointly by The Institute for Genomic Research and VectorBase with support from The Broad Institute of Harvard/MIT.

Release 1.0 has been deposited at GenBank under the accession version AAGE00000000. Long-term curation of the genome sequence and subsequent annotation updates will be the responsibility of VectorBase.

Release 1.0 contains 15,419 high confidence protein-coding genes; alternatively spliced transcripts derived from 992 genes add an additional 1,370 proteins yielding a total of 16,789 predicted proteins. Our structural annotation (gene finding) strategy involved masking repetitive DNA sequences, followed by a combination of EST and protein alignments, trained gene prediction algorithms, and evidence-based gene predictions. Annotation release 1.0 was derived by comparing and merging gene sets generated independently by VectorBase and TIGR.

An autonaming pipeline used sequence similarity to PANTHER, Drosophila (Release 4.3) and UniProt databases as a basis for functional name assignment. Gene Ontology terms were computationally assigned to the genes based on Drosophila GO annotations via sequence homology, or Pfam domain matches above the trusted cutoff. A "supplement" of lower confidence gene predictions is also available at VectorBase.

Aedes MSC and VectorBase Project Plan

The National Institute of Allergy and Infectious Diseases, National Institutes of Health has funded the Aedes aegypti genome project through its Microbial Sequencing Centers (MSCs) at The Institute for Genomic Research (TIGR) and the Broad Institute. The MSCs are responsible for genome sequencing, assembly, and annotation of gene structure and function, with the goal of rapid release of each of these data sets to the scientific community. Once released, the complete sequence and annotation of the Aedes genome will permanently reside at a third NIAID-sponsored entity, VectorBase, which is a Bioinformatics Resource Center (BRC) at the University Of Notre Dame. Delivery of these data will also require coordination with NCBI-GenBank. Given the mutual interests of these organizations, the most effective approach to the initial release of Aedes genomic data will be to work in close collaboration to produce an initial set of annotation, refine and improve the pipelines resident at each of the centers, and to generate data to use in the analysis and publication of Aedes.

We are confident that the combined annotation efforts of the MSCs and VectorBase produce unified, high-quality Aedes annotation for release to the scientific community. This document is a detail project plan intended to describe the complementary activities planned among the three NIAID-funded centers.

See the accompanying document Timeline.pdf for the task timeline overview.

A summary of the activities for the MSC *Aedes* project listed by institution

Institution: Broad

Task	Duration	Start	Finish
Stage 0: Establish Communication Mechanisms Among TIGR, Broad and VectorBase	5 weeks	9/12/05	10/14/05
Develop Assembly Release Strategy	3 weeks	9/12/05	
Conference Call	1 day	9/27/05	
Review and Acceptance of Project Plan	2.4 weeks	9/22/05	
Stage 1: Annotation Preparation	4 weeks	9/19/05	
Release August Assembly	5 days	10/03/05	
Release useful datasets	5 days	10/10/05	
Run and Evaluate Computes	3 weeks	9/26/05	
Stage 2: Production Annotation Gene Structure	8 weeks	10/17/05	

Institution: VectorBase

Task	Dur		
Stage 0: Establish Communication Mechanisms Among TIGR, Broad and VectorBase	5 we		
Conference Call	1 da		
Review and Acceptance of Project Plan	2.4 week		
Stage 1: Annotation Preparation	4 weeks	9/19/05	10/14/05
Release useful datasets	5 days	10/10/05	10/14/05
Stage 2: Production Annotation Gene Structure	8 weeks	10/17/05	12/09/05
VectorBase: gene structure annotation	8 weeks	10/17/05	12/09/05

A summary of the Stages of completion for the MSC *Aedes* project

Stage 0: Establish Communication Mechanisms among TIGR, Broad and VectorBase (5 weeks)

Stage 1: Annotation Preparation (4 weeks)

Stage 2: Production Annotation Gene Structure (8 weeks)

Stage 3: Evaluation of Data (4 weeks)

Stage 4: Data Generation Gene Structure v1.0 (4 weeks)

Stage 5: Production Annotation Functional Computes (3 weeks)

Stage 6: Genbank Submission v1.0 (2 weeks)

Stage 7: Genbank Processing to Release (4 weeks)

Stage 8: Manuscript Preparation (29 weeks)

ID	Task Name	Duration	Start	Finish	Predeces	Notes	Resource Names	Successors
1	**Stage 0: Establish Communication Mechanisms Among TIGR, Broad an**	**25 days**	**Mon 9/12/05**	**Fri 10/14/05**		**ication, such as conference calls or email ...**	**Broad,TIGR,VectorBase**	**12**
2	Develop Assembly Release Strategy	15 days	Mon 9/12/05	Fri 9/30/05		d the primary annotation to the Broad assembly. ...	Broad,TIGR	9
3	Develop Draft Project Plan for Annotation and Analysis	8 days	Mon 9/12/05	Wed 9/21/05		' circulation. The timeframe represented in the d...	TIGR	6,5,11
4	Define Annotation Data types and Metrics for Evaluating Gene Sets	15 days	Mon 9/12/05	Fri 9/30/05		Annotation Data types and File Formats...	TIGR	
5	Conference Call	0 days	Tue 9/27/05	Tue 9/27/05	3	process edits and arrive at a ratified document.	Broad,TIGR,VectorBase	
6	Review and Acceptance of Project Plan	12 days	Thu 9/22/05	Fri 10/7/05	3	eviewed by NIAID and a final version distributed.	Broad,TIGR,VectorBase,NIAID	29
7	**Stage 1: Annotation Preparation**	**20 days**	**Mon 9/19/05**	**Fri 10/14/05**		**uction Annotation will occur at this stage. ...**	**Broad,TIGR,VectorBase**	**12**
8	Create Central Repository (CR)	10 days	Mon 9/19/05	Fri 9/30/05		anged between the Broad, TIGR and Vectorbase.	TIGR	10
9	Release August Assembly	5 days	Mon 10/3/05	Fri 10/7/05	2	m the Broad on a project submission date.</B></d>	Broad	10
10	Release useful datasets	5 days	Mon 10/10/05	Fri 10/14/05	9,8	eleased to the scientific community. The informat...	TIGR,Broad,VectorBase	
11	Define the 10mb region for manual annotation	8 days	Thu 9/22/05	Mon 10/3/05	3	veloped by TIGR and Vectorbase will be identified.	Broad	
12	**Stage 2: Production Annotation Gene Structure**	**40 days?**	**Mon 10/17/05**	**Fri 12/9/05**	**1,7**	**llow us to perform comprehensive evalua...**	**TIGR,Broad,VectorBase**	**29,47**
13	**TIGR gene structure annotation**	**40 days**	**Mon 10/17/05**	**Fri 12/9/05**		**pertain to TIGR's gene structure annotation.**	**TIGR**	
14	Run autopipeline	4 wks	Mon 10/17/05	Fri 11/11/05		g, to the larger and more complex genomes of pl...	TIGR	15
15	Submit TIGR 0.5 in Central Repository	0 days	Fri 11/11/05	Fri 11/11/05	14	vill be deposited in the Central Repository FTP site.	TIGR	16
16	**Iterative improvement of gene set**	**15 days**	**Mon 11/14/05**	**Fri 12/2/05**	**15**	**the individual gene prediction programs. ...**	**TIGR**	
17	Verify against Anopheles and Drosophila	5 days	Mon 11/14/05	Fri 11/18/05		compared to our predicted gene set to identify in...	TIGR	18
18	Locate genes in introns of others	5 days	Mon 11/21/05	Fri 11/25/05	17	nes were identified. Therefore, we will extract ...	TIGR	19
19	EST Data Incorporation	5 days	Mon 11/28/05	Fri 12/2/05	18	EST alignments and the gene structures, and c...	TIGR	20
20	Quality Control	1 wk	Mon 12/5/05	Fri 12/9/05	19	editing of existing gene models, others will resul...	TIGR	21
21	Submit 0.5.1 in Central Repository	0 days	Fri 12/9/05	Fri 12/9/05	20	t will be deposited in Central Repository FTP site.	TIGR	
22	**Vectorbase: gene structure annotation**	**40 days**	**Mon 10/17/05**	**Fri 12/9/05**		**ertebrates and also Anopheles gambiae. ...**	**VectorBase**	
23	autopipeline	4 wks	Mon 10/17/05	Fri 11/11/05		rganisms. There are three main sources genes:...	VectorBase	24
24	make VB v 0.5 available via web	4 wks	Mon 11/14/05	Fri 12/9/05	23	ilable to community for BLAST, SSAHA access....	VectorBase	25
25	Submit 0.5 or 0.5+ into Central Repository	0 days	Fri 12/9/05	Fri 12/9/05	24	will be deposited in Central Repository FTP site....	VectorBase	
26	**Broad: 10 Mb annotation**	**40 days?**	**Mon 10/17/05**	**Fri 12/9/05**		**b). This will provide a complementary dat...**	**Broad**	
27	Annotate 10 Mb region	8 wks?	Mon 10/17/05	Fri 12/9/05		otation for the 10Mb region is requested here..</d>	Broad	28
28	Submit 10 Mb annotation in Central Repository	0 days	Fri 12/9/05	Fri 12/9/05	27	1, will be deposited in Central Repository FTP site.	Broad	
29	**Stage 3: Evaluation of Data**	**20 days?**	**Mon 12/12/05**	**Fri 1/6/06**	**12,6**	**will be deposited into Genbank in Stage 5....**	**Broad,TIGR,VectorBase**	**48,33**
30	implement work in Proposal for evaluating and comparing gene sets	15 days	Mon 12/12/05	Fri 12/30/05		The tasks required for data evaluation include:...	TIGR	31,32
31	define merge strategy based on evaluation	5 days	Mon 1/2/06	Fri 1/6/06	30	decision on the composition of the final gene set.	TIGR,Broad,VectorBase	

Would better BRC-MSC coordination matter?

in support of diagnostics, therapeutics, vaccines, mechanisms, virulence factors etc.

"To call in a statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of." (R.A. Fisher)

- "Faith-based" bottom-up initiatives like the resource centers actually work
- Requirements inputs from expert community and applications developers to MSCs and BRCs to focus efforts and outputs with well-posed problems/questions
- Project management plans to push/pull MSC genomes and MSC generated data and analyses to their corresponding BRCs
- Are MSCs and BRCs separated by a common language and uncommon metrics?
- Collective future of Resource Centers like MSCs and BRCs depends on demonstrating value

Thanks

- NIAID

- TIGR Service Team members of
  - The Joint Technology Center (JTC) – Sequencing
  - Biotechnology Core Services (BCLS & BCIS) – Genome finishing
  - Informatics (IFX) - Assembly, Annotation & Data Submissions

- **Ishwar Chandramouliswaran, MSC Project Manager**